

Predicting graduation in Moroccan open-access bachelors: early indicators and re-enrollment data

Khalid Oqaidi, Sarah Aouhassi, Khalifa Mansouri

Laboratory M2S2I, ENSET of Mohammedia, Hassan II University of Casablanca, Casablanca, Morocco

Article Info

Article history:

Received Apr 11, 2024
Revised Aug 30, 2024
Accepted Sep 28, 2024

Keywords:

Dropout prediction
Educational data mining
Higher education
Machine learning
Open access graduation

ABSTRACT

The primary aim of higher education institutions is the successful graduation of their students. This study explores open-access higher education in Morocco, introducing a predictive model for assessing the probability of students achieving a science bachelor's degree. We analyzed data from 2012 to 2022, initially encompassing 45,573 student entries, and narrowed it down to 14,054 records after data cleaning. Focusing on early academic indicators from enrollment onwards-excluding current program performance—we used popular machine learning classifiers to examine the predictive capacity for student graduation and early dropout. Our comparison included analyses with and without re-enrollment data. Upon analyzing various machine learning algorithms, we attained accuracies between 79% and 86%, identifying random forest (RF) as the superior model for predicting outcomes both with and without incorporating re-enrollment data. This analysis was grounded on initial indicators observed during enrollment and throughout subsequent years, deliberately excluding current academic performance metrics from consideration.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Khalid Oqaidi
Laboratory M2S2I, ENSET of Mohammedia, Hassan II University of Casablanca
Zip Code 28830, Bd Hassan II, Mohammedia, Casablanca, Morocco
Email: khalid.oqaidi@gmail.com

1. INTRODUCTION

The high dropout rates and low graduation rates in Moroccan open-access higher education institutions pose significant challenges to the educational system. With more than 25% of students dropping out after their first year and graduation rates ranging between 27.6% and 33.9%, there is an urgent need for effective interventions to improve student retention and success [1]. This study focuses on predicting student graduation using early enrollment data to enable timely interventions and support for at-risk students.

Early prediction of student attrition in higher education has gained significant attention due to its potential to improve retention rates. Various machine learning techniques, including logistic regression (LR), support vector machines (SVM), decision trees (DT), and artificial neural networks (ANN), have been employed to predict student performance and dropout rates with high accuracy. Key predictors identified in these studies include academic performance, demographic factors, and interaction logs in online learning environments [2]-[6].

Despite these advancements, there is limited research on early prediction using only initial enrollment data, especially in the context of Moroccan open-access institutions. Most existing studies either focus on specific programs or broader university populations in developed countries, often utilizing data collected throughout students' academic careers [2], [7]. Early prediction models have shown effectiveness

even with limited data, such as enrollment variables and first-semester results, highlighting the potential for timely interventions [2], [8].

This study addresses these gaps by focusing on the Moroccan higher education context and utilizing early enrollment data to predict student graduation. Our primary contribution is the comparison of dropout prediction models trained at the moment of the first enrollment using data available at that time, versus models that include additional variables from the second and third re-enrollments. This comparison aims to determine whether significant results can be obtained using only the initial enrollment data, enabling early predictions and interventions before it is too late. This early prediction model is crucial for implementing interventions that can improve student retention and graduation rates in open-access institutions, addressing the alarming statistics of high attrition rates in the first year of university in Morocco.

The paper is structured as follows: in section 2, we provide a comprehensive literature review, discussing recent studies on predictive modeling in education. Section 3 presents the method, detailing the data collection, preprocessing methods, the machine learning models, the experimental setup, and the evaluation metrics used in our study. Section 4 discusses the results, comparing the performance of different models and interpreting the findings. Finally, section 5 concludes the paper by summarizing the key contributions, addressing limitations, and suggesting directions for future research.

2. RELATED LITERATURE REVIEW

The use of machine learning to predict students' performance has been the focus of numerous researchers around the world. The primary concerns of these studies can be categorized into several key areas, which are highly relevant to our study on the early prediction of student attrition in Moroccan higher education. Predicting academic performance: researchers have developed predictive models to forecast students' final results using variables such as age, high school degree score, job status, and country of origin. The researchers [8]-[10] have demonstrated how various demographic and academic factors can influence student outcomes. These studies provide a foundation for understanding which variables might be crucial in predicting student dropout rates early on.

Impact of demographics and behavior on academic success: numerous studies have examined how demographic factors like age, gender, and country of origin influence academic performance. Niyogisubizo *et al.* [4] explored various machine learning algorithms to predict failure using demographic variables. Liao and Wu [5] investigated the impact of digital distractions and demographic features on academic performance. Asthana *et al.* [6] utilized regression-based machine learning models focusing on behavioral and demographic data. These insights are vital for our study as they highlight which demographic factors could be significant predictors of student attrition.

Student retention prediction: predictive models have been developed to determine whether a student is likely to drop out or continue their studies based on demographic information, high school degree scores, and job status. Pek *et al.* [11] demonstrated the effectiveness of machine learning in identifying at-risk students and improving retention through targeted interventions. This aligns with our study's goal of using early enrollment data to predict and mitigate dropout rates.

Residency influence on academic achievement: the impact of students' province of residency and address on their academic performance has been studied, with Ismail and Yusof [12] showing that students from certain regions are more likely to perform better due to the proximity to educational institutions. Understanding the geographic influence is crucial for our context, as regional disparities could affect dropout rates in Moroccan Universities. Comparing high school performance and university performance: studies like those by Mora and Escardíbul [13] have examined the correlation between high school degree scores and university performance to understand how well high school achievements predict academic success at the university level. This is directly relevant to our study, which aims to use early enrollment data, including high school scores, to predict student outcomes.

Effects of job status on academic performance: researchers such as Evangelista [14] have explored how a student's job status impacts their academic achievements, finding differences in performance between working and non-working students. This aspect is relevant to our study as job status might be a significant variable in early prediction models. Gender-based performance analysis: investigations into gender-based differences in academic performance, such as those by [15], [16], provide insights into potential disparities and factors contributing to these differences. Gender is an important demographic factor in our study's predictive models.

International students academic performance: Goller *et al.* [17] analyzed the academic performance of international students, highlighting that local students are more likely to remain in science, technology, engineering, and mathematics (STEM) fields compared to their foreign counterparts. This is pertinent to our research as the presence of international students could impact overall dropout rates. Impact of age on academic achievement: studies by [18], [19] examined how age influences academic performance, finding

that age can significantly impact student success. Including age as a variable in our predictive models can help improve the accuracy of early dropout predictions.

3. METHOD

3.1. Data collection

We gathered three datasets from the Hassan II University of Casablanca Table 1: students: this dataset provides socio-demographic and previous academic features, including student code, country, disability, birth date, gender, city of birth, province of birth, province of residency, baccalaureate (Bac) type, province of bac, honors of Bac, year of Bac, student's socio-professional category (Student SPC), mother's socio-professional category (Mother SPC), father's socio-professional category (Father SPC), and year of enrolment.

Table 1. Description of the features as extracted from the datasets

Datasets category	Original features	Number of values/types	Description
Students	Country	44 codes/integers	The country of nationality
	Disability	8 codes/strings	The kind of disability
	Date of birth	Multiple dates/ 'mm-dd-yy'	
	Gender	2/strings (male, female)	
	City of birth	3964 names/strings	Contains small areas' names too
	Province of birth	134 codes/integers	Bigger than the city of birth
	Province of residency	96 codes/integers	The actual province's address
	Year of enrolment	11 years/integers	Different years of enrollment per student
	Baccalaureate code	90 codes/integers	High school degree type
	Baccalaureate province	98 codes/integers	Where the student gets the baccalaureate
	Baccalaureate honors	5 codes/strings	Abbreviations of honors
	Baccalaureate year	44 years/integers	When the student gets the baccalaureate
	Student's SPC	13 codes/integers	Student's job status
	Mother's SPC	21 codes/integers	Mother's job status
	Father's SPC	36 codes/integers	Father's job status
Results	Degree year	11 years/integers	The year is supposed to pass final exams
	Degree code	66 codes/strings	Degree showing in the results
	Degree version code	19 codes/strings	Groupings of degree codes
	Validation mark	From 0.0 to 20.0 /floats	The mark in the final exam (mark/20)
	Session code	2 codes/1 and 2	Normal and remedial sessions of exams
Enrolment	Result state	5 codes/strings	Result decision about the student after the exam
	Year of enrolment	11 years/integers	Different years of enrollment per student
	Step code	231 codes/strings	Derives from degree code
	Step version code	19 codes/strings	Groupings of degree codes
	Degree code	66 codes/strings	Degree of enrollment
	Degree version code	19 codes/strings	Groupings of degree codes

Degree results: this dataset contains final results for a given degree at enrollment. It includes student code, year of enrolment, degree type, validation version code, validation mark, result state, and session number. Enrolment: this dataset comprises administrative information provided by students upon enrollment. It contains variables such as student code, enrolment year, step code, and degree.

The raw variables, their types, and their possible values are presented in Table 1. As mentioned, there are five different feature categories from the six data categories classified in [20]: academic data before university like baccalaureate honors, academic data inside the university like validation mark, socio-demographic data like Province of residency, financial data like student's spc, and institutional data like degree type. The missing category is the behavioral features set like motivation and attendance.

3.2. Data preprocessing

Data preprocessing is a crucial step in preparing the dataset for machine learning. It involves cleaning and transforming raw data into a format suitable for modeling. Below are the detailed steps taken to preprocess the datasets used in this study.

3.2.1. Student's information dataset

The student's information dataset includes several features that required preprocessing to ensure consistency and suitability for machine learning models:

- Country: there are 44 different countries, coded with integer values between 100 and 521. We re-encoded them in numbers from 1 to 44 in ascending order.

- Disability: there are 8 string disability values ['A', 'M', 'XX', 'V', 'T', 'AV', 'AM', 'MV'], and nan for students without any disability. We re-encoded them with integers from 0 to 8.
- Date of birth: in date format “mm-dd-yy”, we used it with the enrollment year value to extract the student's age at the time of first enrollment.
- Gender: male and female values were encoded with 0 for females and 1 for males.
- City of birth: there are 3,964 different string values, many of which are written in multiple ways. Instead of coding them in integers, we used the province of birth which is more precise.
- Province of birth: there are 134 integer values of the province of birth. We re-encoded them from 1 to 134 in ascending order.
- Province of residency: there are 96 different provinces of residency, coded like “Province of Birth”. We re-encoded them from 1 to 96 in ascending order.
- Year of first enrolment: we did not have explicitly the year of first enrollment in the original dataset but inferred it from the enrollment years for each student.
- Baccalaureate code: there are 90 baccalaureate types coded between 1 and 153. We re-encoded them in ascending order from 1 to 90.
- Baccalaureate Province: 98 provinces of baccalaureate coded between 0 and 117, re-encoded with integers from 1 to 98.
- Baccalaureate honors: we re-encoded the 5 values ('P', 'AB', 'AU', 'B', 'TB') from 1 to 5.
- Baccalaureate year: 44 different values, ranging from 1975 to 2022.
- Student's SPC: 13 values coded with integers between 10 and 99, re-encoded in ascending order from 1 to 13.
- Mother's SPC: 21 values coded with integers between 10 and 99, re-encoded in ascending order from 1 to 21.
- Father's SPC: 36 values coded with integers between 10 and 99, re-encoded in ascending order from 1 to 36.

3.2.2. Enrolment's information dataset

The enrolment's information dataset includes variables that track students' enrollment details over the years:

- Enrollment year: 11 years ranging from 2012 to 2022.
- Degree code: 66 string values, encoded with integers from 1 to 66.
- Degree version code: 19 values coded with integers between 11 and 503, re-encoded with integers from 1 to 19.
- Step code: 231 string codes, re-encoded with integers from 1 to 231.
- Step version code: same codes as degree version code, highly correlated. Only degree version code is considered.

3.2.3. Degree results' dataset

The degree results' dataset includes the final results and other relevant details for each degree:

- Degree year: ranging between 2012 and 2022.
- Degree code: same as degree code in the enrollment dataset.
- Validation mark: float numbers between 0.000 and 20.000, with vacant values for absent students.
- Result state: 4 values ('V', 'ADM', 'AJ', 'NV') coded respectively with integers 1 to 4. Vacant values coded with 0.
- Session code: values 1 for the first session and 2 for the second one, vacant values coded with 0. The target variable 'Graduation' was created from the 'Result State' feature, with values 0 (not graduated) and 1 (graduated). This was verified by comparing with the 'Validation Mark' to ensure correctness.
- Correlation analysis: to ensure that we only use relevant features for our machine learning models, we examined the correlations between different features. The correlation matrix in Figure 1 illustrates the relationships between all features used in the models. Features such as 'Nationality' and 'Disability' showed no significant correlation with 'Graduation' and were excluded from further analysis.

The process of selecting and filtering from the complete dataset to the graduation rate of the targeted cohorts is presented in Figure 2. We considered only the cohorts making their first enrollment in 2013, 2014, 2015, and 2016, to track the graduation results from 2016 to 2022. We can point out that the bachelor's degree students represent the majority with 39,628 out of 45,573 (86.95%), illustrating the claim about open-access institutions. The tracked cohorts making their first matriculation in 2013, 2014, 2015, and 2016, are supposed to graduate normally in 2016, 2017, 2018, and 2019 respectively. We extend the observation to 2020, 2021, and 2022. The ending graduation rate was 29.6%. This aligns closely with previous reports, which tracked open-access bachelors from three Moroccan universities matriculated between 2007 and 2013, concluding a graduation rate between 27.6% and 34.4%.

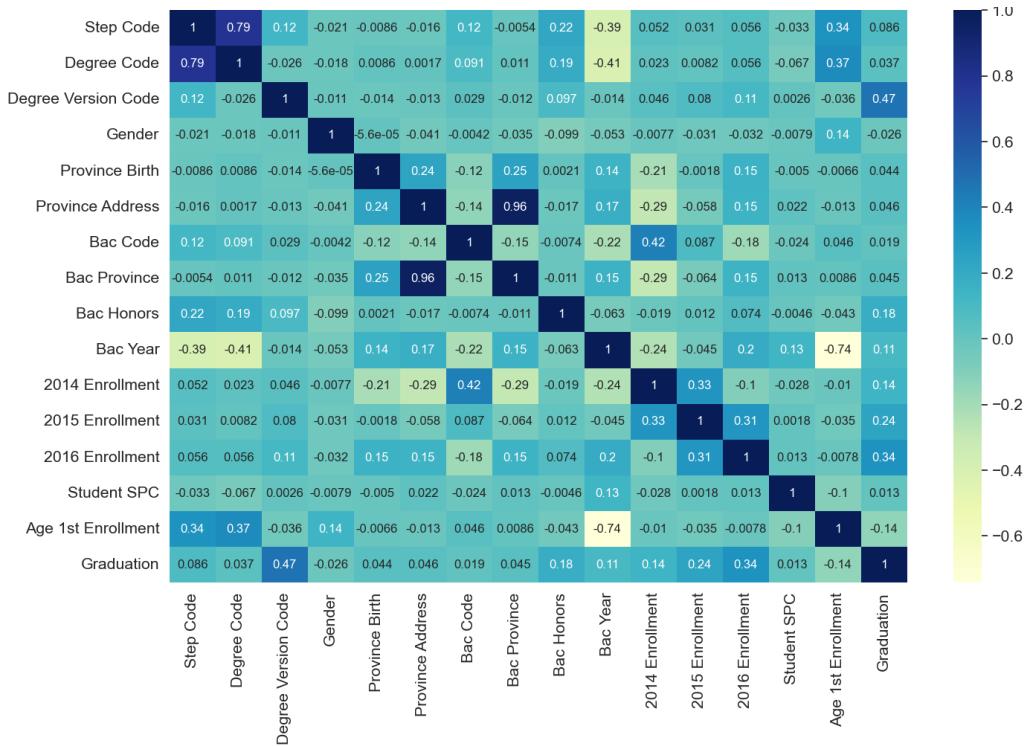


Figure 1. Correlation matrix between all features used in the machine learning models

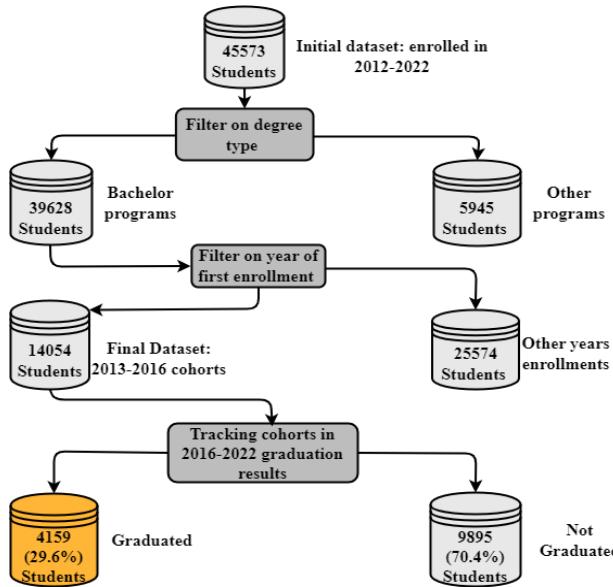


Figure 2. 2013-2016 sciences bachelor degrees' cohorts' graduation rate

3.3. Machine learning models

To evaluate the impact of early enrollment data on graduation prediction, we trained machine learning algorithms using two sets of independent variables:

- List 1: ['Step Code', 'Degree Version Code', 'Gender', 'Province Birth', 'Province Address', 'Bac Code', 'Bac Honors', 'Bac Year', '2014 Enrollment', '2015 Enrollment', '2016 Enrollment', 'Student SPC', 'Age 1st Enrollment'].
- List 2: ['Step Code', 'Degree Version Code', 'Gender', 'Province Birth', 'Province Address', 'Bac Code', 'Bac Honors', 'Bac Year', 'Student SPC', 'Age 1st Enrollment'].

The first set of variables includes enrollment year features ('2014 Enrollment', '2015 Enrollment', '2016 Enrollment') to capture information about subsequent enrolments. The second set excludes these features to test if predictions can be accurately made using only the data available at the time of first enrollment. If the models trained with list 2 achieve satisfactory results, it would be significant as it would enable early prediction of graduation outcomes, allowing interventions before the student even starts their courses.

We employed the following machine learning algorithms, selected based on their effectiveness as demonstrated in previous studies [20], [21]:

- LR: widely used for binary classification problems, LR predicts the probability of a sample belonging to a certain class. It has been proven effective in student attrition prediction [22].
- DT: suitable for both binary and multi-class classification problems, DT have shown efficiency in predicting dropout rates in higher education contexts such as engineering courses [23].
- RF: this algorithm combines multiple DT to improve accuracy. It is the most frequently used dropout prediction algorithm according to a systematic review [24].
- SVM: known for its performance in both binary and multi-class classification, SVM has shown predictive efficiency alongside LR [25].
- kNN: kNN is simple yet effective for both binary and multi-class classification. Over 13% of dropout prediction studies in a systematic review employed this algorithm [24].
- Training strategy: we trained each algorithm twice: once with list 1 and once with list 2. This approach allowed us to compare the models' predictive performance with and without the enrollment year features. The primary objective was to determine whether graduation predictions could be reliably made using only the data available at first enrollment, thus enabling early intervention.

3.4. Evaluation metrics

We used several evaluation metrics to assess the performance of the machine learning models:

- Confusion matrix: comprising true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) Table 2.

Table 2. General confusion matrix

Actual positive	TP	FN type I error
Actual negative	FP	TN
	Type II error	
	Predicted positive	Predicted negative

- Accuracy: the proportion of correct predictions (TP+TN) divided by the total number of instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

- Precision: the ratio of TP to the sum of TP and FP.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

- Recall: the ratio of TP to the sum of TP and FN.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

- F1-score: the harmonic mean of recall and precision, balancing the two metrics.

$$\text{F1 - Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

These metrics, especially in binary classification, provide a comprehensive assessment of model performance [26].

4. RESULTS AND DISCUSSION

4.1. Results summary

In this section, we present the results of the five machine learning algorithms mentioned earlier. We experimented with different train/test dataset ratios: 90%/10%, 80%/20%, and 70%/30% as shown in Table 3. We selected the ratio with the best accuracy, and when the accuracy was similar, we chose the ratio

with the lowest FN value. For some algorithms, a 90% train/10% test split was optimal, while for others, a 70% train/30% test split was preferable.

For each algorithm, we trained the model with independent variables from lists 1 and 2, representing the presence ('yes') and absence ('no') of re-enrollment features (2014, 2015, and 2016). Each result involved six experiments:

$$NE = NR \times NVL \quad (5)$$

where: NE is number of experiments; NR is number of ratios; and NVL is number of variable lists.

The results table contains the values of TP, TN, FP, FN, precision, recall, accuracy, and F1-score. TP values are the correctly predicted graduates, while FN values are the incorrectly predicted non-graduates. Our primary objective is to accurately predict students at risk of not graduating.

Table 3. Prediction results for five ML models with/without re-enrollment features

ML algorithm	LR	DT	RF	kNN	SVM
Complete features	yes	no	yes	no	yes
Training	70%	90%	70%	70%	70%
Test	30%	10%	30%	30%	30%
TP	556	177	809	856	876
TN	2875	971	2631	2664	2720
FP	81	30	325	292	236
FN	705	228	452	405	385
Precision	0.8207	0.8229	0.8115	0.8239	0.8496
Recall	0.8041	0.8165	0.8157	0.8136	0.8527
Accuracy	0.8041	0.8165	0.8157	0.8136	0.8527
F1-score	0.7766	0.7951	0.8128	0.7919	0.8499
					0.8543
					0.8150
					0.8415
					0.7684
					0.7807

As we use weighted averaging, the recall and accuracy have the same results. The best F1-score and recall were 0.854 and 0.857 respectively, registered in the random forest (RF_algorithm with and without the three additional features. The best precision was 0.850 registered in RF with the additional features used. The best value of FN was 371 in RF with (70% train, 30% test), while in kNN we found the smallest value of FN 140 with (90% train, 10% test). The best value of FP (the minimum value) was 32 for (70% train, 30% test) in SVM with additional features used, and 12 for (90% train, 10% test) in the same algorithm without additional features used. The best TN (the maximum value) was 2875 with (70% train, 30% test) in LR, and 989 with (90% train, 10% test) in SVM. The best TP (the maximum value) was 890 in RF with (70% train, 30% test), and 265 with (90% train, 10% test).

4.2. Critical analysis

RF: the RF algorithm demonstrated the highest performance, achieving F1-scores of 0.854 (with re-enrollment features) and 0.857 (without re-enrollment features). This suggests that RF is particularly adept at handling the complexity of our data, making it a reliable model for predicting student graduation outcomes.

kNN: the kNN algorithm also showed strong results, particularly with the 90% train/10% test split, achieving an FN of 140 and an F1-score of 0.845. This indicates kNN's effectiveness, especially when a larger training set is available.

SVM: the SVM algorithm had the best FP rate, indicating it is less likely to incorrectly predict a student will graduate when they will not. This is crucial for minimizing unnecessary interventions and optimizing resource allocation.

LR and DT: both algorithms demonstrated moderate performance but were outperformed by RF and kNN in most metrics. LR and DT are still valuable for their simplicity and interpretability, but they may require additional tuning or feature engineering to match the performance of more complex models.

- Interpretation of results: the findings indicate that while all models benefit from the inclusion of re-enrollment features, the impact is less significant than expected. The high performance of models trained solely on early enrollment data (list 2) suggests that timely and accurate predictions can be made as early as the first enrollment, enabling early interventions.
- Comparison with existing studies: these results align with previous studies that highlight the effectiveness of RF and kNN in educational data mining. For instance, a study on clustering students' admission data using k-means, hierarchical, and DBSCAN algorithms found that data preparation and clustering methods are critical for educational data analysis. Additionally, the same study demonstrated the effectiveness of ensemble learning techniques, such as RF, in improving student performance

prediction. The consistency of our findings with existing literature reinforces the robustness of these algorithms for dropout and graduation predictions [27].

5. CONCLUSION

This study developed and evaluated machine learning models to predict student graduation outcomes in Moroccan open-access universities using early enrollment data. The findings highlight the effectiveness of RF and kNN algorithms, with both achieving high accuracy and F1-scores. Notably, models that utilized only initial enrollment data performed comparably to those incorporating re-enrollment features, demonstrating the potential for early prediction and timely intervention to improve student retention. By identifying RF as the most effective model, followed closely by kNN, this research underscores the capability of early enrollment data to drive proactive support measures in educational institutions. However, the study's limitations include the exclusion of behavioral data, such as student engagement and attendance, which could further enhance predictive accuracy, and the focus on a single institution, which may impact the generalizability of the findings. Strict ethical guidelines were followed to protect student privacy, with sensitive data anonymized to ensure integrity and compliance. Future research should aim to incorporate additional data, such as financial status and academic engagement metrics, to provide a more comprehensive view of the factors influencing student graduation. Expanding the study to other faculties and institutions would help validate the model's generalizability, while longitudinal analyses could offer valuable insights into the long-term effects of early interventions on student outcomes. This work contributes to the growing body of evidence supporting the use of machine learning for early identification of at-risk students, ultimately aiding in the development of targeted strategies to enhance student success and reduce dropout rates.

ACKNOWLEDGEMENTS

The authors thank Hassan II University of Casablanca for the data used in this study.

REFERENCES

- [1] B. Rahma *et al.*, *Higher education in Morocco: effectiveness, efficiency and challenges of the open access university system (in French: L'enseignement supérieur au Maroc: efficacité, efficience et défis du système universitaire à accès ouvert)*, Conseil Supérieur de l'Éducation, de la Formation et de la Recherche Scientifique, 2018.
- [2] M. Segura, J. Mello, and A. Hernández, "Machine learning prediction of university student dropout: does preference play a key role?," *Mathematics*, vol. 10, no. 18, pp. 1–20, 2022, doi: 10.3390/math10183359.
- [3] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interactive Learning Environments*, vol. 30, no. 8, pp. 1414–1433, 2022, doi: 10.1080/10494820.2020.1727529.
- [4] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, no. March, p. 100066, 2022, doi: 10.1016/j.caai.2022.100066.
- [5] C. H. Liao and J. Y. Wu, "Deploying multimodal learning analysis models to explore the impact of digital distraction and peer learning on student performance," *Computers and Education*, vol. 190, 2022, doi: 10.1016/j.compedu.2022.104599.
- [6] P. Asthana, S. Mishra, N. Gupta, M. Derawi, and A. Kumar, "Prediction of student's performance with learning coefficients using regression based machine learning models," *IEEE Access*, vol. 11, pp. 72732–72742, 2023, doi: 10.1109/ACCESS.2023.3294700.
- [7] M. Barramuño, C. Meza-Narváez, and G. Gálvez-García, "Prediction of student attrition risk using machine learning," *Journal of Applied Research in Higher Education*, vol. 14, no. 3, pp. 974–986, May 2022, doi: 10.1108/JARHE-02-2021-0073.
- [8] E. Alhazmi and A. Sheneamer, "Early predicting of students performance in higher education," *IEEE Access*, vol. 11, pp. 27579–27589, 2023, doi: 10.1109/ACCESS.2023.3250702.
- [9] S. A. Priyambada, T. Usagawa, and M. ER, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge," *Computers and Education: Artificial Intelligence*, vol. 5, 2023, doi: 10.1016/j.caai.2023.100149.
- [10] M. A. Baig, S. A. Shaikh, K. K. Khatri, M. A. Shaikh, M. Z. Khan, and M. A. Rauf, "Prediction of students performance level using integrated approach of ML algorithms," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 01, pp. 216–234, Jan. 2023, doi: 10.3991/ijet.v18i01.35339.
- [11] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, and R. Alhajj, "The role of machine learning in identifying students at-risk and minimizing failure," *IEEE Access*, vol. 11, no. January, pp. 1224–1243, 2023, doi: 10.1109/ACCESS.2022.3232984.
- [12] N. Ismail and U. K. Yusof, "A systematic literature review: recent techniques of predicting STEM stream students," *Computers and Education: Artificial Intelligence*, vol. 5, 2023, doi: 10.1016/j.caai.2023.100141.
- [13] T. Mora and J. O. Escardibul, "Schooling effects on undergraduate performance: evidence from the University of Barcelona," *Higher Education*, vol. 56, no. 5, pp. 519–532, 2008, doi: 10.1007/s10734-007-9108-y.
- [14] E. Evangelista, "An optimized bagging ensemble learning approach using BESTrees for predicting students' performance," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 10, pp. 150–165, 2023, doi: 10.3991/ijet.v18i10.38115.
- [15] B. Nita *et al.*, "Machine learning in the enrolment management process: a case study of using GANs in postgraduate students' structure prediction," *Procedia Computer Science*, vol. 207, no. Kes, pp. 1350–1359, 2022, doi: 10.1016/j.procs.2022.09.191.
- [16] Y. Safsouf, K. Mansouri, and F. Poirier, "Tabat: Design and experimentation of a learning analysis dashboard for teachers and learners," *Journal of Information Technology Education: Research*, vol. 20, pp. 331–350, 2021, doi: 10.28945/4820.
- [17] D. Goller, A. Diem, and S. C. Wolter, "Sitting next to a dropout: academic success of students with more educated peers," *Economics of Education Review*, vol. 93, 2023, doi: 10.1016/j.econedurev.2023.102372.
- [18] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V. Moniaga, "Determining factors that affect student performance using

various machine learning methods," *Procedia Computer Science*, vol. 216, no. 2022, pp. 597–603, 2022, doi: 10.1016/j.procs.2022.12.174.

[19] A. Rafique *et al.*, "Integrating learning analytics and collaborative learning for improving student's academic performance," *IEEE Access*, vol. 9, pp. 167812–167826, 2021, doi: 10.1109/ACCESS.2021.3135309.

[20] K. Oqaidi, S. Aouhassi, and K. Mansouri, "Towards a students' dropout prediction model in higher education institutions using machine learning algorithms," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 18, pp. 103–117, 2022.

[21] S. D. A. Bujang *et al.*, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[22] M. Phan, A. D. Caigny, and K. Coussement, "A decision support framework to incorporate textual data for early student dropout prediction in higher education," *Decision Support Systems*, vol. 168, 2023, doi: 10.1016/j.dss.2023.113940.

[23] A. M. Mariano, A. B. D. M. L. Ferreira, M. R. Santos, M. L. Castilho, and A. C. F. L. C. Bastos, "Decision trees for predicting dropout in Engineering Course students in Brazil," *Procedia Computer Science*, vol. 214, pp. 1113–1120, 2022, doi: 10.1016/j.procs.2022.11.285.

[24] D. Andrade-Girón *et al.*, "Predicting student dropout based on machine learning and deep learning: a systematic review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 5, pp. 1–11, 2023, doi: 10.4108/eetsis.3586.

[25] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, no. November 2018, p. 106189, 2020, doi: 10.1016/j.chb.2019.106189.

[26] J. P. Bernius, S. Krusche, and B. Bruegge, "Machine learning based feedback on textual student answers in large courses," *Computers and Education: Artificial Intelligence*, vol. 3, no. March, p. 100081, 2022, doi: 10.1016/j.caai.2022.100081.

[27] E. L. Cahapin, B. A. Malabag, C. S. Santiago Jr., J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, Dec. 2023, doi: 10.11591/eei.v12i6.4849.

BIOGRAPHIES OF AUTHORS



Khalid Oqaidi holds a Ph.D. in Computer Science and Artificial Intelligence from the Faculty of Science and Technology (FST) of Mohammedia, Hassan II University of Casablanca. His research focuses on using data science techniques to improve higher education quality. He is also a Telecommunications Engineer from the National Institute of Posts and Telecommunications (INPT Rabat). He is the founder and CEO of Entscheider, a consulting firm specializing in statistical data analysis, prediction using machine learning and data science techniques, and market research. He can be contacted at email: khalid.oqaidi@gmail.com.



Sarah Aouhassi is an Information Systems and Data Quality Professor at ENSAM Engineering School of Casablanca, she holds a Doctorate in Information Systems from the Faculty of Sciences of Casablanca, and a State Engineering degree in applied statistics from the National Institute of Statistics and Applied Economy (INSEA Rabat). She has ten years of experience as a statistical and assessment chief officer at Hassan II University of Casablanca. Her research is focused on information systems quality and its interaction with higher education quality. She can be contacted at email: aouhassi.sarah@gmail.com.



Khalifa Mansouri is a Computer Science Professor and researcher at the Hassan II University of Casablanca, ENSET of Mohammedia. Real-time systems, information systems, e-learning systems, and industrial systems (modeling, optimization, and numerical computing) are the main areas of his research. He obtained a Ph.D. in Calculation and Optimization of Structures from Mohammed V University in Rabat in 1994, an HDR in 2010, and a Ph.D. in Computer Science from Hassan II University of Casablanca in 2016. He can be contacted at email: khalifa.mansouri@enset-media.ac.ma.